Bayesian Model Diagnostics Based on Artificial Autoregressive Errors

M. Peruggia, The Ohio State University

Technical Report No. 758

July, 2005

Department of Statistics
The Ohio State University
1958 Neil Avenue
Columbus, OH 43210-1247
Bayesian Model Diagnostics Based on Artificial Autoregressive Errors

Mario Peruggia

Department of Statistics
The Ohio State University
Columbus, OH 43210-1247

July 2005

Author’s Footnote

Mario Peruggia is Associate Professor, Department of Statistics, The Ohio State University, Columbus, OH 43210. This material is based upon work supported by the National Science Foundation under Awards No. SES-021457, and SES-0437251. The author thanks Peter Craigmile, Steve MacEachern and Trisha Van Zandt for their helpful comments.
Abstract

Hierarchical Bayes models provide a natural way of incorporating covariate information into the inferential process through the elaboration of regression equations for one or more of the model parameters, with errors that are often assumed to be i.i.d. Gaussian. Unfortunately, building adequate regression models is a complicated art form that requires the practitioner to make numerous decisions along the way. Assessing the validity of the modeling decisions is often difficult.

In this article I develop a simple and effective device for ascertaining the quality of the modeling choices and detecting lack-of-fit. I specify an artificial autoregressive structure (AAR) in the probability model for the errors that incorporates the i.i.d. model as a special case. Lack-of-fit can be detected by examining the posterior distribution of AAR parameters. In general, posterior distributions that assign considerable mass to a region of the AAR parameter space away from zero provide evidence that apparent dependencies in the errors are compensating for misspecifications of some other aspects (typically conditional means) of the model. I illustrate the methodology through several examples including its application to the analysis of data on brain and body weights of mammalian species and response time data.

Keywords: Allometry; Asymptotic normality; Autocorrelation; Hierarchical models; Response times.

1 Introduction

It is common statistical practice to evaluate the fit of a model by examining the behavior of the realized residuals. For example, in multiple linear regression analyses
(Weisberg 1985), the residuals are often plotted against one of the covariates or against the predicted values of the corresponding cases. The plots are then scanned for the presence of non-random patterns that would call into question the assumption of an i.i.d. error structure. Realized residuals incompatible with i.i.d. errors may indicate inadequacies in the model and suggest avenues to improve the fit (e.g., transformations of the response and/or some of the predictors, introduction of new predictors, removal of old predictors, etc.). The use of residual plots and related diagnostics (added variables plots, q-q plots, etc.) is well established in the frequentist arena and virtually all statistical software packages provide easy access to computational and graphical tools to perform residual-based model checks.

Hierarchical Bayes models (Carlin and Louis 2000, Gelman, Carlin, Stern, and Rubin 2004) provide a natural way of incorporating covariate information into the inferential process through the elaboration of regression equations for one or more of the model parameters, with errors that are often assumed to be i.i.d. Gaussian. The approach is conceptually simple and the development of efficient Markov Chain Monte Carlo (MCMC) computational techniques has made its implementation feasible even in rather complicated settings, especially since flexible and user-friendly computational environments such as BUGS (Spiegelhalter, Thomas, Best, and Gilks 1996a) and WinBUGS (Spiegelhalter, Thomas, Best, and Lunn 2003) have become available. Unfortunately, building adequate regression models is a complicated art form that requires the practitioner to make numerous hard decisions along the way. By comparison to its frequentist counterpart, the contents of the Bayesian model-building toolbox look quite scanty. All the more so when the regression model to be constructed is for a parameter at a higher stage in the hierarchy—a quantity for which, typically, the intuition is not as well developed as the intuition for quantities
that are directly observable.

In this article I demonstrate that residual analysis is also a powerful diagnostic tool for Bayesian model building. Rather than examining the realized residuals directly (a difficult task, especially when the regression models are for parameters that appear at higher levels of the model hierarchy), I introduce an artificial autoregressive (AAR) structure in the probability model for the errors that incorporates the i.i.d. model as a special case. Lack-of-fit can be detected by examining the posterior distribution of the AAR parameters. In general, posterior distributions that assign considerable mass to a region of the AAR parameter space away from zero provide evidence that apparent dependencies in the errors are compensating for misspecifications of some other aspects (typically conditional means) of the model.

Econometricians have long recognized that, given a data vector \( Y \) whose elements are collected sequentially over time, the misspecification of the design matrix \( X \) in a linear regression model \( Y = \theta X + e \) can induce autocorrelation in the elements of \( e \) (Judge, Griffiths, Hill, and Lee 1980). Within this context, the Durbin and Watson test is the earliest and most popular frequentist technique for detecting autocorrelation in the residuals (Durbin and Watson 1950 and 1951). More recently, a Bayesian comparison of regression models with i.i.d. and AR(1) errors was presented in Carota (1998). In econometric applications the misspecification of the design matrix is typically difficult to remedy, due in part to the fact that some of the important regressors may not be known or measurable. Thus, econometricians end up retaining the misspecified \( X \) and are principally concerned with the impact that the presence of autocorrelation has on the estimation of \( \theta \) and how to optimally estimate \( \theta \) in its presence. Bayesian estimation in this context is considered, for example, by Zellner and Tiao (1964) using locally uniformly priors on the model parameters and in...
Chapter 5 of Judge et al. (1980) using Jeffreys’ prior.

The basic premise that model misspecification leads to autocorrelated errors also motivates the development of the diagnostic tools presented in this article. However, the perspective here is essentially opposite than in econometrics: the aim of the diagnostics is to assess lack-of-fit, uncover specific shortcomings of the model, and suggest possible directions for further elaboration and improvement that would make a model with conditionally i.i.d. errors appropriate. The applicability of the AAR error structure to uncover lack-of-fit in the context of elaborate Bayesian hierarchical models is very flexible and can be readily implemented using popular MCMC simulation software. While the AAR structure implies that the observations are ordered, I need not require that they be collected over time. As the examples presented throughout the article will make clear, in certain cases the ordering will be indeed determined by the time when the observations arose, in others it will be suggested naturally by the specific application, and in others yet it will be entirely arbitrary.

The remainder of the article is organized as follows. Section 2 builds some intuition for the technique by considering an elementary example and deriving some asymptotic properties. Sections 3 and 4 evaluate the performance of the technique in a variety of settings using simulated and real data respectively. Finally, Section 5 contains a brief discussion and some concluding remarks.

2 Intuition and Asymptotic Considerations

In this section, I consider a simple model for which direct closed-form calculation can be performed and present some asymptotic results to develop intuition and motivate the development of model diagnostics based on the specification of AAR error
structures.

**Example 2.1** I am planning to collect \( n \) i.i.d. observations, \( y_i \), from a \( N(\mu_y, \sigma_y^2) \) distribution with known variance \( \sigma_y^2 \) and to make inference about the unknown mean \( \mu_y \). My friend Pythia claims she has the ability to make inferences without seeing the data and tells me that the value of the mean is equal to some fixed real number \( a \). Pythia has a spotty track-record of reporting the right mean and I wish to verify her assertion by actually collecting the data and analyzing them. To do so, always assuming conditional independence unless stated otherwise, I specify the following model with AAR errors for the \( y_i \). For \( i = 1, \ldots, n \), let

\[
y_i = a + \eta_i,
\]

where \( a \) is a fixed constant, and \( \eta_1 \sim N(0, \sigma_1^2) \). Furthermore, for \( i = 2, \ldots, n \), let

\[
\eta_i = \phi \eta_{i-1} + \epsilon_i,
\]

with \( \epsilon_i \sim N(0, \sigma^2) \) and \( \phi \sim N(0, \sigma_\phi^2) \), where \( \sigma_1^2, \sigma^2, \) and \( \sigma_\phi^2 \) fixed and known. The model has conjugate structure, and the posterior distribution of \( \phi \) is easily seen to be normal with mean and variance

\[
E(\phi|y) = \frac{\sum_{i=2}^{n}(y_i - a)(y_{i-1} - a)}{(\sigma_\phi^2/\sigma_\epsilon^2) + \sum_{i=2}^{n}(y_{i-1} - a)^2} \quad \text{and} \quad V(\phi|y) = \left[ \frac{1}{\sigma_\phi^2} + \frac{1}{\sigma_\epsilon^2} \sum_{i=2}^{n}(y_{i-1} - a)^2 \right]^{-1}.
\]

Suppose the true mean of the \( y_i \) is \( \mu_y = 0 \). Then, denoting by \( \rho_y \) the lag-1 autocor-
relation of the \( y_i \) (which is zero because of the independence assumption),

\[
\lim_{n \to \infty} \frac{1}{(n-1)} \sum_{i=2}^{n} (y_i - a)(y_{i-1} - a) = \\
= \lim_{n \to \infty} \left[ \frac{1}{(n-1)} \sum_{i=2}^{n} y_i y_{i-1} - \frac{a}{(n-1)} \sum_{i=2}^{n} y_i - \frac{a}{(n-1)} \sum_{i=2}^{n} y_{i-1} + \frac{a^2}{(n-1)} (n-1) \right] \\
= \rho_y - a \mu_y - a \mu_y + a^2 = a^2,
\]

and

\[
\lim_{n \to \infty} \frac{1}{(n-1)} \sum_{i=2}^{n} (y_{i-1} - a)^2 = \\
= \lim_{n \to \infty} \left[ \frac{1}{(n-1)} \sum_{i=2}^{n} y_i^2 - \frac{2a}{(n-1)} \sum_{i=2}^{n} y_{i-1} + \frac{a^2}{(n-1)} (n-1) \right] \\
= \sigma_y^2 - 2a \mu_y + a^2 = \sigma_y^2 + a^2.
\]

Thus, \( \lim_{n \to \infty} E(\phi|y) = \frac{a^2}{(a^2 + \sigma_y^2)} \) and \( \lim_{n \to \infty} (n-1)V(\phi|y) = \sigma_y^2/(a^2 + \sigma_y^2) \).

The asymptotic posterior mean of \( \phi \) is zero if Pythia is right (i.e., when \( a = 0 \)). The further she is from the truth (i.e., the larger \( a^2 \) is), the closer the asymptotic posterior mean of \( \phi \) will be to one. The asymptotic variance, in turn, is determined by the relative sizes of \( \sigma_y^2 \) and \( a^2 + \sigma_y^2 \). These considerations suggest that draws from the posterior distribution of \( \phi \) can be used to diagnose Pythia’s accuracy. A histogram of posterior \( \phi \) values centered somewhere in the neighborhood of zero would give no indication that her guess should be questioned.

From a more general perspective, let \( f \) and \( g \) be smooth functions on a finite closed interval, that, without loss of generality, can be taken to be \([0, 1]\). I will consider regular infill asymptotics, assuming that the true data generating mechanism is given by \( y_i = f(i/n) + u_i \), where, for \( i = 0, \ldots, n \), the \( u_i \) are independent normal errors.
with mean zero and variance $\sigma_u^2$. The fitted model with AAR errors is given by
\[ y_i = g(i/n) + \eta_i, \]
with $\eta_i = \phi \eta_{i-1} + \epsilon_i$, where the i.i.d. zero-mean normal noise process $\{\epsilon_i\}$ has variance $\sigma^2_\epsilon$ and is independent of the true noise process $\{u_i\}$.

Let $r(t) = f(t) - g(t)$ denote the bias at the point $t$ induced by fitting the wrong regression function $g$. Simple algebraic manipulations yield $[r(i/n) + u_i] = \phi [r((i-1)/n) + u_{i-1}] + \epsilon_i$. Treating the process $z_i = [r(i/n) + u_i]$ as zero-mean stationary, $\phi$ is estimated consistently for any fixed $n$ by
\[
\hat{\phi}_n = \frac{(n+1)^{-1} \sum_{i=1}^n z_i z_{i-1}}{(n+1)^{-1} \sum_{i=0}^{n-1} (z_i)^2}.
\]

As $n$ goes to infinity, the numerator of Equation (1) converges to the limit of
\[
\sum_{i=1}^n \frac{r(i/n)r((i-1)/n)}{n+1} + \sum_{i=1}^n \frac{r(i/n)u_{i-1}}{n+1} + \sum_{i=1}^n \frac{r((i-1)/n)u_i}{n+1} + \sum_{i=1}^n \frac{u_i u_{i-1}}{n+1}.
\]
By the definition of a Riemann sum, the first term in the summation converges to $\int_0^1 r^2(t) \, dt$. Also, because of the smoothness assumptions on $f$ and $g$ and of the various distributional assumptions, the SLLN implies that the last three terms converge to zero. Hence the numerator converges to $\int_0^1 r^2(t) \, dt$. Similarly, the denominator converges to the limit of
\[
\sum_{i=0}^n \frac{r^2(i/n)}{n+1} + \frac{2 \sum_{i=0}^n r(i/n)u_i}{n+1} + \frac{\sum_{i=0}^n u_i^2}{n+1}.
\]
The first term in the summation converges to $\int_0^1 r^2(t) \, dt$, the second term converges to zero, and the last term converges to $\sigma_u^2$. Hence the denominator converges to $\int_0^1 r^2(t) \, dt + \sigma_u^2$. Combining the results for the numerator and the denominator shows
that \( \hat{\phi}_n \) converges with probability one to

\[
\frac{\int_0^1 r^2(t) \, dt}{\int_0^1 r^2(t) \, dt + \sigma_u^2}.
\]  

(2)

The expression in Equation (2) can be interpreted as a *signal to noise ratio* for detecting unmodeled trend: the integrated squared bias in the numerator quantifies the amount of unmodeled trend and the denominator quantifies the amount of unmodeled trend plus the amount of noise in the data.

In practice \( g(t) \) is unknown. A full Bayesian model would typically specify a parametric family \( g(t|\theta) \) to which \( g(t) \) belongs as well as a prior distribution for the model parameters \( \theta, \phi, \) and \( \sigma^2 \). For any fixed \( n \), the model will define a likelihood \( p_n(y_n|\theta, \phi, \sigma^2) \) for the data \( y_n = (y_0, \ldots, y_n) \) collected under the regular infill scheme. Let \( (\theta^*, \phi^*, (\sigma^*_e)^2)_n \) be the value of \( (\theta, \phi, \sigma^2) \) that minimizes the Kullback-Leibler divergence of \( p_n(y_n|\theta, \phi, \sigma^2) \) from the true distribution of the data, \( p_n(y_n) \), let \( (\theta^*, \phi^*, (\sigma^*_e)^2) = \lim_{n \to \infty} (\theta^*, \phi^*, (\sigma^*_e)^2)_n \), and let \( V(\theta, \phi, \sigma^2) = \lim_{n \to \infty} [(n + 1)I_n(\theta, \phi, \sigma^2)]^{-1} \), where \( I_n(\theta, \phi, \sigma^2) \) denotes the Fisher information matrix based on a sample of size \( n + 1 \). Under suitable regularity conditions (see Gelman, Carlin, Stern, and Rubin (2004) for the case of i.i.d. observations), \( \sqrt{n+1} \) times the joint posterior distribution of \( (\theta, \phi, \sigma^2) \) will converge to a normal distribution with mean \( (\theta^*, \phi^*, (\sigma^*_e)^2) \) and covariance matrix \( V[(\theta^*, \phi^*, (\sigma^*_e)^2)] \). Marginally, \( \sqrt{n+1} \) times the posterior distribution of \( \phi \) will converge to a univariate normal distribution with mean

\[
\phi^* = \frac{\int_0^1 [r^*(t)]^2 \, dt}{\int_0^1 [r^*(t)]^2 \, dt + \sigma_u^2}
\]  

(3)

and variance \( (\sigma^*_e)^2/(\int_0^1 [r^*(t)]^2 \, dt + \sigma_u^2) \). The expression for the mean \( \phi^* \) is the same as expression (2) (and can be interpreted similarly), except that \( r(t) \) is replaced by
$r^*(t) = f(t) - g(t|\theta^*)$, where $g(t|\theta^*)$ is the member of the family $g(t|\theta)$ that best approximates $f(t)$ in the sense of minimizing $\int_0^1 (f(t) - g(t|\theta))^2 dt$. An illustration of this asymptotic approximation is presented in Example 3.1.

3 Examples Using Simulated Data

In this section, I examine the finite sample and asymptotic performance of the AAR diagnostic device by presenting some simulated data examples. To explain the intuition behind the methodology, the examples in this section deal with simple models for which more direct diagnostics can be readily developed. In several cases the posterior distributions of the AAR parameters could be derived using low dimensional numerical integration, but I chose to estimate them by MCMC simulation to reflect the way in which the diagnostic device would typically be used for general hierarchical Bayes models.

Example 3.1 This example illustrates the methodology in the context of a linear regression model. The data set is comprised of $n = 101$ pairs of observations $(x_i, y_i)$ simulated from the model $y_i = 1.0 + (0.1)x_i + u_i$, where, for $i = 0, \ldots, n$, $x_i = i$ and the $u_i$ are independent standard normal errors.

Using BUGS, I fit several regression models with an AAR error structure and a possibly misspecified slope. For $i = 0, \ldots, n$, I assume that $y_i = \alpha + bx_i + \eta_i$, where $b$ is a fixed constant and $\alpha \sim N(0, 10^3)$. The errors $\eta_i$ have the same basic AAR structure as in Example 2.1, but now the variances $\sigma_1^2$ and $\sigma_2^2$ are given inverse gamma prior distributions. Specifically, $\eta_1 \sim N(0, \sigma_1^2)$, and, for $i = 1, \ldots, n$, $\eta_i = \phi \eta_{i-1} + \epsilon_i$, with $\epsilon_i \sim N(0, \sigma_\epsilon^2)$, $1/\sigma_1^2 \sim Gamma(2.5, 10.0)$, $1/\sigma_2^2 \sim Gamma(0.1, 0.1)$, and $\phi \sim N(0, 1.0)$.
Figure 1 displays the data and five fitted regression lines corresponding to five specifications of the slope $b$ in the regression equation. The five values of $b$ under consideration are $b_1 = 0.10$, $b_2 = 0.09$, $b_3 = 0.07$, $b_4 = 0.05$, and $b_5 = 0.00$. For $1 \leq j \leq 5$, the equation of the fitted regression line is given by $y = \hat{\alpha}_j + b_jx$, where $\hat{\alpha}_j$ is an estimate of the posterior mean of $\alpha$ obtained by averaging 1,000 posterior draws generated by BUGS based on the model with slope $b_j$. In each of the five cases, I discarded the first 10,000 draws and subsampled every tenth of the subsequent 10,000 ones. Denoting by $\bar{y}$ the sample mean of the observed $y$ values, each estimated intercept $\hat{\alpha}_j$ arises from the shrinkage of the value $\bar{y} - b_j \times 50$ towards the prior mean of zero.

As evidenced by the figure, the horizontal line (corresponding to slope $b_5$) provides the worst fit. As $x$ increases, the residuals from the fitted line progressively shift from having large negative values to having large positive values (see the bottom left panel of Figure 2). This increasing trend is due to the erroneous specification of the conditional mean for the observations. Failure to remove the trend induces strong positive autocorrelations in the residuals that are captured by the AAR model. As a consequence, the posterior distribution of the AAR parameter $\phi$, illustrated by the histogram in the bottom right panel of Figure 2, is concentrated near one. Based on the 1,000 draws from the posterior distribution of $\phi$ used to construct the histogram, the posterior mean is estimated at 0.93 and the 95% equal-tailed posterior probability interval is $[0.83, 1.01])$.

The best fitting line appears to be the one rising most rapidly (corresponding to slope $b_1 = 0.1$). This, of course, is not surprising because the data were simulated from a regression model with slope equal to 0.1. In this case, a model of independence for the AAR error structure ($\phi = 0$) would be appropriate as evidenced by the
absence of trends in the residual plot in the top left panel of Figure 2. Accordingly, the posterior distribution of the AAR parameter $\phi$, illustrated by the histogram in the top right panel of Figure 2, is centered in the vicinity of zero. Based on the 1,000 draws from the posterior distribution of $\phi$ used to construct the histogram, the posterior mean is estimated at 0.02 and the 95% equal-tailed posterior probability interval is $[-0.19, 0.24]$. The intermediate cases corresponding to slopes $b_2$, $b_3$, and $b_4$, are characterized by correspondingly worsening fits (see the left intermediate panels of Figure 2) and posterior distributions of the AAR parameter $\phi$ that become progressively more concentrated about a central location that shifts from zero towards one (see the right intermediate panels of Figure 2).

The asymptotic posterior mean $\phi^*$ of $\phi$ under regular infill can be computed using Equation (3), adjusting for the fact that the the unit interval in the original derivation of Section 2 is here replaced by an interval of length 100. Note that the initialization of the AAR process with the specification of the variance of $\eta_1$ is not considered explicitly in the derivations of Section 2 because it does not impact on the asymptotic arguments. Observing that $\int_0^{100} (1 - 0.10t - (\alpha + bt))^2 dt$ is minimized at $\alpha^* = 6 - 50b$, yields $100^{-1} \int_0^{100} [r^*(t)]^2 dt = (10b - 1)^2 (25/3)$ and

$$\phi^* = \frac{(25/3)(10b - 1)^2}{(25/3)(10b - 1)^2 + 1}, \quad (4)$$

with $(25/3)(10b - 1)^2$ capturing the amount of unmodeled trend and $\sigma^2_u = 1$ capturing the amount of noise in the data. Figure 3 illustrates the convergence of the posterior distribution of $\phi$ to normality for the case of $b_4 = 0.05$, The three panels display, from bottom to top, histograms of posterior draws of $\phi$ based on samples of sizes 101, 1,001, and 10,001, respectively. The histogram become more symmetric and concentrate
more closely around the asymptotic mean value of 0.676 calculated according to Equation (4). In each panel, the value of the asymptotic mean is represented by the vertical segment and the sample mean of the posterior draws is represented by a triangular symbol.

**Example 3.2** The AAR diagnostic performs equally well when the true conditional mean of the observations is a polynomial of a higher degree than the one specified by the model. In this example, the data were simulated as \( y_i = (0.002)x_i^2 + \epsilon_i \), where, for \( i = -50, \ldots, 50 \), \( x_i = i \) and the \( \epsilon_i \) are independent standard normal errors.

A first model states that \( y_i = \alpha + \beta x_i + \gamma x_i^2 + \eta_i \), where \( \alpha, \beta, \) and \( \gamma \sim N(0, 10^3) \) independently. The rest of the model is specified as in Example 3.1. A second model incorrectly drops the quadratic term \( \gamma x_i^2 \) in the specification of the conditional mean of the observations, while leaving all other specifications unchanged.

The two models yield two fitted regression curves. The first is the parabola \( y = \hat{\alpha} + \hat{\beta} x + \hat{\gamma} x^2 \) obtained by averaging 1,000 posterior draws of \( \alpha, \beta \) and \( \gamma \) generated according to the first model by running BUGS with a burn-in of 10,000 and a thinning rate of 1 in 10. The second is the straight line obtained in a similar way on the basis of the second model’s incorrect specification of the functional form of the conditional mean. The residuals from the fitted curves behave differently. In particular, the incorrect specification of the degree of the polynomial for the conditional mean gives rise to a pronounced quadratic trend in the residuals of the second model and, in turn, the AAR error structure tries to compensate for the presence of this trend.

This is reflected in the posterior distribution of the AAR parameter \( \phi \). The posterior mean of \( \phi \) corresponding to the first model specification is estimated at 0.01 with 95% equal-tailed posterior probability interval given by \([-0.20, 0.20]\). For the incorrect specification of the second model, the estimated posterior mean of 0.71 is much
larger and the 95% equal-tailed posterior probability interval given by [0.55, 0.88] is slightly narrower. This is the same qualitative behavior observed in Example 3.1.

**Example 3.3** In this example, building on Examples 3.1 and 3.2, I assess the performance of the AAR diagnostic in the context of a linear repeated measures model. I simulated repeated measurements $y_{ij}$ according to the model 

$$y_{ij} = a_i + b_i x_j + c x_j^2 + u_{i;j},$$

where $i, 1 \leq i \leq 10$, can be thought of as indexing an experimental subject and $j, 1 \leq j \leq 20$, indexes a measurement taken at $x_j = j$. For the sake of discussion, I can think of the $x_j$ as representing successive time points at which the measurements were taken. For $1 \leq i \leq 9$, I sampled independently $a_i$ from a $N(3, (0.5)^2)$ distribution and $b_i$ from a $N(0.2, (0.05)^2)$ distribution, and set $c_i = 0$. Furthermore, I set $a_{10} = 2.26787$, $b_{10} = 0.44321$, and $c_{10} = -0.01108$. Thus, the regression curves for the first nine subjects are linear with expected measurements of 3.2 at $x_1 = 1$ and 7.0 at $x_{20} = 20$. The regression curve for the last subject is quadratic, takes on values 2.7 at $x_1 = 1$ and 6.7 at $x_{20} = 20$, and has a horizontal tangent at $x_{20} = 20$. I generated the measurement errors $u_{ij}$ from a $N(0, (0.2)^2)$ distribution. The longitudinal regression curves and corresponding simulated data are depicted in the two panels of Figure 4, where the thinner lines represent the curves corresponding to the last subject.

The correct assessment of the functional form of the regression curves for the various subjects is of paramount importance in the analysis of repeated measures data. Often, the regression curves are modeled as low degree polynomials. Naturally the question arises of whether a straight line is appropriate or a higher degree polynomial is needed to improve the fit. With this question in mind, I fit the model 

$$y_{ij} = \alpha_i + \beta_i x_j + \eta_{i,j},$$

where the subject specific regression parameters are independently distributed as $\alpha_i \sim N(\alpha_0, \sigma_\alpha^2)$, $\beta_i \sim N(\beta_0, \sigma_\beta^2)$, $1 \leq i \leq 10$, with $\alpha_0, \beta_0 \sim N(0, 10^3)$ and $1/\sigma_\alpha^2, 1/\sigma_\beta^2 \sim Gamma(0.1, 0.1)$, all independently.
For each subject, the error terms $\eta_{i,j}$ follow the same AAR structure introduced in Example 3.1. Thus, there is an AAR parameter $\phi_i$ corresponding to each of the ten subjects. The posterior distributions of the ten AAR parameters are summarized by the histograms of Figure 5, based on 1,000 MCMC draws generated by BUGS with a burn-in of 10,000 and a thinning rate of 1 in 10. While the posterior distributions of the first nine AAR parameters do not appear to indicate any substantial lack-of-fit, the posterior mass of $\phi_{10}$ is concentrated well to the right of 0, with a posterior mean estimated at 0.75. The AAR diagnostic tool is thus very sensitive to the inadequacy in the specification of the regression curve for subject 10.

4 Examples Using Real Data

Example 4.1 This example demonstrates that the AAR method can also be employed when the independent variable $x$ does not play the role of time and does not take on equally spaced values. For illustration, I consider the 100 complete pairs of brain and body weights for placentat mammalian species published in Sacher and Staffeldt (1974). In addition to the weights, the data set also records the order and sub-order to which each species belongs. This data set was used in MacEachern and Peruggia (2002) to illustrate the performance of some numerical and graphical diagnostic tools in detecting the shortcomings of a simple linear regression (SLR) model compared to a variance component (VC) model. For mammal $i$, I denote by $y_i$ the natural logarithm of the brain weight and by $x_i$ the logarithm of the body weight recentered about the sample mean of the log body weights. A scatterplot of the data (not shown) suggests that a SLR model should provide an excellent fit to the data.

The Bayesian specification of the SLR model with ARR errors says that, for
\( i = 1, \ldots, 100, \ y_i = \alpha + \beta x_i + \eta_i, \) where, independently, \( \alpha \sim N(0, 10^4) \) and \( \beta \sim N(2/3, (2/9)^2) \) (see MacEachern and Peruggia 2002 for a detailed justification of these prior specifications). The prior distribution of the AAR error structure follows the same specifications given in Example 3.1.

The Bayesian VC model with AAR errors incorporates random effect terms \( \gamma_\ell \) and \( \delta_m \) for the 13 orders and 19 sub-orders in the taxonomy by saying that, for \( i = 1, \ldots, 100, \ y_i = X'_i \theta + \eta_i, \) where \( \theta = (\alpha, \beta, \gamma_1, \ldots, \gamma_{13}, \delta_1, \ldots, \delta_{19})' \) and \( X \) is the \( 100 \times (2 + 13 + 19 = 34) \) design matrix constructed by adjoining a column of ones, the column \( x \) of recentered log body weights, 13 columns of 0-1 order indicators, and 19 columns of 0-1 sub-order indicators. The priors for \( \alpha \) and \( \beta \) are specified as in the SLR model and, independently, \( \gamma_\ell \sim N(0, \sigma_\gamma^2), \ \ell = 1, \ldots, 13, \ \delta_m \sim N(0, \sigma_\delta^2), \ m = 1, \ldots, 19, \ 1/\sigma_\gamma^2 \sim Gamma(0.1, 0.1), \) and \( 1/\sigma_\delta^2 \sim Gamma(0.1, 0.1). \) Once again, the prior distribution of the AAR error structure is as in Example 3.1.

The histograms of Figure 6, based on 1,000 draws generated using BUGS with a burn-in of 10,000 and a thinning rate of 1 in 10, summarize the posterior distributions of the AAR parameter \( \phi \) for the SLR and VC models. For the SLR model, the posterior expectation of \( \phi \) is estimated at 0.61 with a 95% equal-tailed posterior probability interval given by [0.45, 0.78]. For the VC model, the posterior expectation of \( \phi \) is estimated at 0.28 with a 95% equal-tailed posterior probability interval given by [-0.07, 0.59].

There is a clear indication that the SLR model is not adequately capturing the variability in the data despite the strong linear trend evidenced by the scatterplot. The reasons for this are carefully outlined in MacEachern and Peruggia (2002). Essentially, the SLR model neglects to account for dependencies of species within orders and sub-orders. The VC model ameliorates the situation by inducing positive cor-
relations within these groups. In this example, the AAR diagnostic device exploits the existence of a trade-off between the residual autocorrelation structure and the presence of random effects in the model (cf. Pinheiro and Bates 2000, p. 398). Specifically, when the correlations between groups are not taken into consideration by the SLR model, the residuals within each order and sub-order tend to behave similarly and the AAR error structure is sensitive to this grouping because the observation are arranged in the data file according to orders and sub-orders.

One might wonder what role the arrangement of the species within each order and sub-order plays in determining the strength of the results. Figure 7 helps to address this question. All three panels display the residuals from the least squares fit of the SLR model. What differs is the order in which the residuals are arranged along the horizontal axis. In all three panels the species sub-orders share the same arrangement (Marsupialia, Insectivora, Chiroptera, Prosimii, Anthropoidea, Edentata, Lagomorpha, Sciuromorpha, Myomorpha, Hystricomorpha, Cetacea, Fissipeda, Pinnipedia, Proboscidea, Hyracoidea, Perissodactyla, Suiformes, Tylopoda, Ruminantia). In addition, the middle panel maintains the species arrangement of the original data file. In the top panel the species are rearranged in such a way that, within sub-order, the residuals come along in increasing order, so as to maximize autocorrelation. In the bottom panel, the arrangement of the species within each sub-order follows a random permutation. Visual inspection confirms that the autocorrelation is highest in the top panel. The autocorrelation also appears to be larger in the middle panel than in the bottom one. In fact, the values of the lag one sample autocorrelations for the arrangements in the three panel are 0.6574, 0.5863, and 0.4839 from top to bottom and one would expect the posterior estimates of the AAR parameter $\phi$ in the Bayesian SLR model to follow a similar pattern.
As it turns out, with the observations arranged as in the top panel, the posterior distribution of $\phi$ was estimated on the basis of an MCMC run of size 1,000 as having a mean of 0.68 with a 95% equal-tailed posterior probability interval given by [0.51, 0.83]. The arrangement depicted in the bottom panel yielded a posterior mean for $\phi$ of 0.54 and a 95% equal-tailed posterior probability interval of [0.35, 0.72], also estimated on the basis of an MCMC run of size 1,000. These results confirm the intuition developed in the preceding paragraph that the posterior distribution of $\phi$ for the arrangement in the top panel is stochastically larger than the posterior corresponding to the the original arrangement depicted in the middle panel (posterior mean of 0.61 and 95% posterior probability interval of [0.45, 0.78], as seen earlier), which in turn is stochastically larger than the posterior corresponding to the random arrangement depicted in the bottom panel.

There are $1! \times 4! \times 1! \times 3! \times 18! \times 1! \times 2! \times 4! \times 11! \times 9! \times 3! \times 14! \times 4! \times 1! \times 1! \times 2! \times 4! \times 2! \times 15!$ or approximately $10^{60}$ permutations of the species within sub-order that preserve a fixed arrangement of the sub-orders. In addition to the permutation depicted in the bottom panel of Figure 7, I generated another 99 random permutations and fit the Bayesian SLR model with AAR error structure. The 100 estimates of the posterior expectation of $\phi$ ranged from 0.47 to 0.70 with a mean value of 0.62 which is very close to the estimate of 0.61 obtained from the original data file arrangement. All 100 estimates are well above the value of 0.28 obtained from the fit of the VC model. This indicates that, regardless of the arrangement of the various species within sub-orders, the AAR diagnostic device is capable of signaling the inadequacies of the SLR model.

I also considered arrangements that would only keep the taxonomy orders fixed, allowing for permutation of the species to run across sub-orders. Such arrangements typically yield lower estimated values of the posterior mean of the AAR parameter.
For 100 randomly generated arrangements of this type the smallest estimate of the posterior expectation of $\phi$ was 0.36 and the largest was 0.61 with a mean of 0.51. Here, all 100 estimated posterior expectations were smaller than the estimate of 0.61 obtained from the original data file arrangement but, again, they were all larger than the estimate of 0.28 obtained from the VC model. Thus, grouping by taxonomic order alone would still induce dependencies that are strong enough to enable the AAR diagnostic device to detect the limitations of the SLR model compared to the VC model.

Example 4.2 In this example I look at a data set of Williams (1959). This data set was analyzed by Carlin and Chib (1995) and Spiegelhalter, Thomas, Best, and Gilks (1996b) to illustrate the use of Bayes factors to compare non-nested regression models. The response variable $y$ is the maximum compressive strength parallel to the grain measured on 42 specimens of radiata pine. There are two possible predictors: $x$, denoting the density of the specimens, and $z$, denoting the density adjusted for resin content. Corresponding to the two predictors, there are two competing SLR models with AAR errors that are specified, for $i = 1 \ldots 42$, as

Model A: $y_i = \alpha + \beta x_i + \eta_i$ and Model B: $y_i = \gamma + \delta z_i + \eta_i$.

Carlin and Chib (1995) and Spiegelhalter et al. (1996b) use different transformations of the variables in their analyses with independent errors. Carlin and Chib (1995) recenter the $x_i$ and the $z_i$ by subtracting off their means while Spiegelhalter et al. (1996b) standardize them by subtracting off their means and dividing by their standard deviations and they also similarly standardize the response values $y_i$. Slightly different priors are also used. Yet, both articles report Bayes factors that demonstrate
Table 1: Posterior Estimates of $\phi$ For the Pines Data Example.

<table>
<thead>
<tr>
<th>Ordering</th>
<th>Model A Expectation</th>
<th>95% Prob. Int.</th>
<th>Model B Expectation</th>
<th>95% Prob. Int.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>-0.02</td>
<td>(-0.35, 0.32)</td>
<td>-0.15</td>
<td>(-0.49, 0.20)</td>
</tr>
<tr>
<td>Based on $x - z$</td>
<td>0.35</td>
<td>(0.00, 0.72)</td>
<td>-0.01</td>
<td>(-0.35, 0.32)</td>
</tr>
<tr>
<td>Based on Residuals</td>
<td>0.26</td>
<td>(-0.10, 0.63)</td>
<td>-0.13</td>
<td>(-0.46, 0.19)</td>
</tr>
</tbody>
</table>

unequivocally how Model B provides a better fit than Model A.

I will now show how the AAR error diagnostic device is able to detect the superiority of Model B over Model A. In my analyses I used the same standardized variables $x$, $y$ and $z$ and the same prior specification as in Spiegelhalter et al. (1996b). In particular, I set $\alpha \sim N(0, 10^{-6})$, $\beta \sim N(0, 10^{-4})$, $\gamma \sim N(0, 10^{-6})$, and $\delta \sim N(0, 10^{-4})$, independently within models, with the AAR error structure following the basic specification of Example 3.1, except that the shape and scale parameters of the gamma distribution for $1/\sigma^2_\epsilon$ were set equal to $10^{-4}$.

In this example there is no natural ordering of the observations, not even a partial ordering, such as that induced by the taxonomy of the mammals data set of Example 4.1. If one fits Models A or B using the original ordering in the data file, the posterior distribution of the AAR parameter $\phi$ turns out to put considerable mass around zero and no evidence of lack-of-fit is uncovered in either case. The posterior distributions of $\phi$ corresponding to the two fits are summarized in the first line of Table 1.

Here, however, the main concern is a comparison of the fit provided by the two models and the ordering of the observations should relate to this comparison. The models differ in the predictor they use, so the relative quality of the fit will be determined by differences in the two predictors. This suggest simply ordering the data according to the increasing values of the differences $x_i - z_i$ (alternatively, an ordering based on relative differences, normalized for the size of the predictors, could also be
The posterior distributions of $\phi$ corresponding to the two fits of the ordered data are summarized in the second line of Table 1. While the posterior distribution of $\phi$ in Model B continues to put considerable mass around zero, the posterior distribution of $\phi$ in Model A is shifted to the right of zero, thus providing evidence of lack-of-fit. Model A fits poorly systematically at those points where the differences between the $x$ and $y$ predictors are large, a feature that AAR diagnostic can readily detect.

A similar behavior of the posterior distributions of $\phi$ can be observed when the observations are ordered according to the average differences of the residuals for the models with independent errors (or, equivalently, according to the average differences of the fitted values). As a convenient shortcut, because of the noninformative nature of the prior distributions, the observations can be ordered according to the differences of the residuals from the least squares fits. This approach yields the posterior distributions for $\phi$ summarized in the third line of Table 1. These summaries still indicate that Model B provides a superior fit, albeit not as strongly as the summaries in the second line of the table. The ordering based on the differences of the residuals is particularly useful in situations when the models being compared contain several, possibly different, predictors.

**Example 4.3** In this example I examine the ability of the AAR diagnostic device to uncover aspects of model inadequacy in the context of a complex hierarchical model for a large set of response time data. The data were collected in a series of recognition memory trials conducted on four subjects over ten non-consecutive days.

For each trial, a subject was initially asked to study a list of 40 words randomly selected from a database of 2337 common English words. The subject was then presented with a sequence of 40 words, 20 selected from the study list and 20 selected
from words in the database not included in the study list. The words were displayed sequentially on a computer monitor and the subject was asked to strike one of two keys on the keyboard depending on whether she thought the word was an “old” word contained in the study list or a “new” word not contained in the study list. The times in milliseconds elapsed between the appearances on the words on the screen and the keystrokes were recorded, along with an indicator of the accuracy of the responses. Each subject participated in two trials on each day. There were thus a total of $4 \times 2 \times 10 = 80$ trials, each contributing 40 response times.

The hierarchical model considered here is a refinement of the model presented in Peruggia, Van Zandt, and Chen (2002). For the current analysis I used a two parameter Weibull likelihood to model the shifted response times obtained by subtracting a value equal to 95% of the the minimum response time for each list from all of the 40 response times for that list. (A three parameter Weibull likelihood could be used to model the unshifted response times.) Let $RT_{i,d,l,w}$ denote the shifted response times, where $i, 1 \leq i \leq 4,$ indexes subject, $d, 1 \leq d \leq 10,$ indexes day, $l, 1 \leq l \leq 2,$ indexes list, and $w, 1 \leq w \leq 40,$ indexes word. Conditional on $r_{i,d,l}$ and $\lambda_{i,d,l,w},$ the $RT_{i,d,l,w}$ are assumed to follow independently a Weibull distribution with shape parameter $r_{i,d,l}$ and scale parameter $\lambda_{i,d,l,w},$ for $1 \leq w \leq 40.$

The logarithm of the scale parameters is endowed with a regression structure that includes random effects $\alpha_{i,d,l}$ primarily intended to model different levels of subject specific learning as days go by. The regression also includes fixed effects for the nature of the words (“old” vs. “new”), the accuracy of the responses (“right” vs. “wrong”), and their interaction. In summary, I assumed that $\ln(\lambda_{i,d,l,w}) = \alpha_{i,d,l} + \beta_1 I.0ld_{i,d,l,w} + \beta_2 I.Right_{i,d,l,w} + \beta_3 (I.0ld_{i,d,l,w} \times I.Right_{i,d,l,w}) + \eta_{i,d,l,w},$ with $I.0ld$ denoting a 0-1 indicator of an old word and $I.Right$ denoting a 0-1 indicator of a
correct response. For the random effects I assumed \( \alpha_{i,d,l} \sim N(\mu_d, \sigma^2) \), independently, where, also independently, \( \mu_d \sim N(0, 10^{-4}) \) and \( 1/\sigma^2 \sim \text{Gamma}(10^{-1}, 10^{-1}) \). For the fixed effects coefficients I assumed \( \beta_1, \beta_2, \beta_3 \sim N(0, 10^{-3}) \), independently. For the shape parameters I assumed that, independently, \( r_{i,d,l} \sim \text{Exp}(\theta_d) \) where, also independently, \( \theta_d \sim \text{Exp}(\theta_0) \) and \( \theta_0 \sim \text{Exp}(10^{-3}) \).

The error terms \( \eta_{i,d,l,w} \) act as diagnostic devices for the stated model of conditional independence of the 40 response times within each list, following the basic AAR structure used throughout the article with AAR coefficients \( \phi_{i,d,l} \) that depend on subject, day, and list. In this example, I specified a more diffuse prior for the variance of the first innovation than in the previous examples by assuming \( 1/\sigma^2 \sim \text{Gamma}(0.1, 0.1) \) rather than \( 1/\sigma^2 \sim \text{Gamma}(2.5, 10.0) \). I made this choice because, with 80 different lists, the data can contribute much information about the distribution of the first innovation. The AAR structure is specified with respect to a given ordering of the response times within a list. In the diagnostic analysis that follows I will consider two separate orderings, the one corresponding to the sequence in which the words were presented to the subjects and the one corresponding to a rearrangement in which the “new” words are made to precede the “old” words, while preserving the original ordering within each group.

For a basic diagnostic analysis, I quickly scanned the histograms of the 80 sets of 1,000 realizations from the marginal posterior distributions of the AAR parameters \( \phi_{i,d,l} \) generated using WinBUGS with a burn-in period of 20,000 and a subsequent thinning rate of 1 in 500. Most distributions put considerable mass around zero but there are several interesting cases in which the shapes of the posterior distributions of the AAR parameters suggest inadequacies in the fit for certain word lists. For the purpose of illustration I will now look at some of these cases.
The histograms of Figure 8 represent estimates of the posterior densities of \( \phi_{1,2,1} \) and \( \phi_{3,9,1} \). The posterior distribution of \( \phi_{3,9,1} \) puts 0.97 of its mass to the right of zero and has a mean of 0.67. The posterior distribution of \( \phi_{1,2,1} \) shows similar qualitative features (if not quantitatively as pronounced), putting 0.91 of its mass to the right of zero and having a mean of 0.50. To understand what causes the mass of the posterior distributions of the two autoregressive parameters to be so heavily shifted to the right of zero, I computed residuals from the Bayesian fit as follows. With the superscript \((k)\), \(1 \leq (k) \leq 1,000\), indexing the \( k \)-th MCMC parameter draw, let 
\[
\lambda^{(k)}_{i,d,l,w} = \exp \left[ \alpha^{(k)}_{i,d,l} + \beta^{(k)}_{1} I.\text{Old}_{i,d,l,w} + \beta^{(k)}_{2} I.\text{Right}_{i,d,l,w} + \beta^{(k)}_{3} (I.\text{Old}_{i,d,l,w} \times I.\text{Right}_{i,d,l,w}) \right]
\]
and let \( E \left( W^{(k)}_{i,d,l,w} \right) \) denote the mean of a Weibull random variable with shape parameter \( r^{(k)}_{i,d,l} \) and scale parameter \( \lambda^{(k)}_{i,d,l,w} \). The residual for the response time \( RT_{i,d,l,w} \) is then computed as 
\[
\text{res}_{i,d,l,w} = RT_{i,d,l,w} - (1,000)^{-1} \sum_{k=1}^{1,000} E \left( W^{(k)}_{i,d,l,w} \right).
\]
Note that these are observation-level residuals which, although related, do not correspond directly to the AAR error terms appearing at a higher level of the hierarchy.

The two panels of Figure 9 display the residuals for the response times of subject 1, day 2, list 1 and subject 3, day 9, list 1. After adjusting for the values of the covariates, the earlier response times for subject 3, day 9, list 1 appear to be slower than the model would predict (the residuals are mainly positive) and the later response times appear to be faster (the residuals are mainly negative). This systematic departure from a random pattern in the residuals is accurately detected by the AAR parameter diagnostic. The residuals for subject 1, day 2, list 1 also appear to exhibit a systematic pattern, corresponding to four or five alternating stretches of longer and shorter response times. The pattern in the set of residuals for subject 3 appears to be stronger than the pattern in the set of residual for subject 1 and this difference is reflected in the fact that the posterior distribution of \( \phi_{3,9,1} \) is shifted further to the
right than the posterior distribution of $\phi_{1,2,1}$.

Another interesting case is the one illustrated in Figures 10 and 11, dealing with subject 3, list 6, day 1. The left panel of Figures 10 summarizes the posterior distribution of $\phi_{3,6,1}$ when the RTs are ordered according to the sequence in which the words were originally presented. The histogram is essentially symmetric about zero and provides no indication of lack-of-fit. Contrast this to the right panel which summarizes the posterior distribution of $\phi_{3,6,1}$ when the RTs are rearranged so that those corresponding to the “new” words are made to precede those corresponding to the “old” words, while preserving the original ordering within each group. Now the posterior mass is heavily shifted to the right of zero. Inspection of the residual plots of Figure 11 provides an explanation for this difference. The top panel, corresponding to the original ordering, does not present any systematic pattern, but the bottom panel suggests that responses to the new words were systematically faster than the model would predict and responses to the old words were systematically slower.

In this example the AAR error devices provide a simple and reliable means of scanning the 80 word lists to examine the fit of the model. The examination of the histograms of the posterior MCMC draws of the AAR parameters is much quicker and easier to interpret than a direct examination of the observation-level residuals from the model fit. Once lack-of-fit is suspected for a certain list a more thorough analysis, including a direct examination of the residuals, can pinpoint more specifically the nature of the inadequacies in the model. In addition to those described above, there are a variety of additional model deficiencies and uncharacteristic observations that the diagnostic device is able to uncover. For example, there are a few cases in which the distribution of the AAR parameter is shifted to the left of zero, corresponding to a systematic pattern of the residuals that alternate rhythmically between positive
and negative values.

The complexity of these data makes modeling a challenge. Overall the specified model is adequate, but there are a few lists for which the fit is not entirely satisfactory. Even though the individual experimental tasks are relatively short (there are only 40 words in each list) and long term trends and dependencies are not the norm, in some cases the assumption of conditional independence is seemingly violated. For example, it appears as if subject 3, in her reaction to the words in list 1 on day 9, is trying to compensate for a slow start in the first half of the list by speeding up her responses to the words in the second half of the list. This and other types of uncommon features cannot be captured by the basic hierarchical model, but the introduction of the AAR error structure acts as an effective screening tool for uncovering them.

5 Discussion

In this article I have proposed a general procedure based on AAR errors for detecting lack-of-fit in hierarchical Bayes model. The principal appeal of the procedure lies in its flexibility and its ease of implementation. All examples presented in this article were in fact fit using publicly available software without any need to write customized programs.

The most popular Bayesian tools for performing model comparison and assessing goodness-of-fit—such as, for example, Bayes factors (Kass and Raftery 1995) and their generalizations (Berger and Pericchi 1996; O’Hagan 1995), posterior predictive model checks (Gelman, Meng, and Stern 1996), and the nice, general-purpose method based on a Bayesian version of the $\chi^2$ test for goodness-of-fit recently proposed in Johnson (2004)—do not examine the residuals directly. The behavior of the residuals, however,
can shed much light on the fit of a model and suggest directions for refinement if lack-of-fit is detected. The strength of the proposed method is that the AAR device can be applied to assess the behavior of residuals at different levels of the hierarchy, probing different aspects of the fit, without much conceptual or practical difficulty.

Unlike other diagnostics for model fit based on embedding (such as the decision theoretical model elaboration strategy of Carota, Parmigiani, and Polson 1996 and its application to the detection of autocorrelation of the disturbances in regression models described in Carota 1998), there is no expectation that the model providing the embedding (the model with AAR errors in this case) be adequate. The sole purpose of the AAR error device is to detect if, and in what ways, the embedded model with independent errors is inadequate. In this respect, even though more elaborate models of dependence could be considered, the first-order autoregressive structure is all that is needed because it is sensitive to misspecifications of conditional means and is easy to understand. In my experience, initialization of the first error term $\eta_1$, though arbitrary, does not have a great impact on the inferential conclusions, as long as the prior distribution of the variance $\sigma_1^2$ is not too diffuse.

The potential benefits of the method when applied to complex hierarchical models is clearly illustrated in Example 4.3. There, I introduced, at a higher level of the hierarchy, a separate AAR error device for each of 80 sets of multidimensional observations (the 80 sets of RTs for each of the 80 word lists). The posterior distributions of each AAR parameter is a low dimensional summary that can quickly detect if the model of conditional independence provides an inadequate fit to the RTs for the corresponding list. Scanning such summaries to assess and quantify lack-of-fit (cf. Figures 8 and 10) is much easier than scanning observation-level residual plots (cf. Figures 9 and 11).
The AAR device is based on orderings of the observations. To detect lack-of-fit, as illustrated in the various examples, it is important not to find an ordering for which the AAR parameter is significant (after all there will always be such an ordering) but to find a meaningful ordering for which this is true. While a routine implementation of the method should always consider orderings based on the size of predicted values and of covariate values, other orderings might be suggested by application-specific considerations.

References


Figure 1: Simulated Data and Five Fitted Lines for the Linear Regression Model of Example 3.1.

Figure 2: Residuals from the Five Fitted Lines and Corresponding Histograms of the AAR Parameters for the Linear Regression Model of Example 3.1. The triangular symbols represent the estimated posterior means of the AAR parameters.
Figure 3: Convergence to Normality Under Regular Infill of the Posterior Distributions of the AAR Parameters in the Linear Regression Model of Example 3.1 for the Case $b_4 = 0.05$. The vertical segments represent the value of the asymptotic mean and the triangular symbols represent the estimated posterior means of the AAR parameters.
Figure 4: Simulated Longitudinal Regression Curves (left panel) and Data (right panel) for Example 3.3. The thicker lines correspond to the last subject who has a quadratic regression.

Figure 5: Histograms of the AAR Parameters for Example 3.3. The triangular symbols represent the estimated posterior means of the AAR parameters.
Figure 6: Histograms of the AAR Parameters for the Two Models for the Mammals Data Fit in Example 4.1. The triangular symbols represent the estimated posterior means of the AAR parameters.

Figure 7: Three Different Orderings of the Residuals from the Least Squares Fit of the SLR Model to the Mammals Data of Example 4.1.
Figure 8: Histograms of the AAR Parameters for Two Word Lists in Example 4.3. The triangular symbols represent the estimated posterior means of the AAR parameters.

Subject 1; Day 2; Original List 1

Subject 3; Day 9; Original List 1

Figure 9: Residuals for the Response Times based on the Two Word Lists Considered in Figure 8. A symbol ‘N’ denotes a “new” word and a symbol ‘O’ denotes an “old” word.
Figure 10: Histograms of the AAR Parameters for Two Orderings of a Word List in Example 4.3. The triangular symbols represent the estimated posterior means of the AAR parameters.

Figure 11: Residuals for the Response Times based on the Two Orderings of the Word List Considered in Figure 10. A symbol ‘N’ denotes a “new” word and a symbol ‘O’ denotes an “old” word.